PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Effectiveness of serious games and impact of design elements on engagement and educational outcomes in healthcare professionals and students: a systematic review and meta-analysis protocol.
AUTHORS	Maheu-Cadotte, Marc-André; Cossette, Sylvie; Dubé, Veronique; Fontaine, Guillaume; Mailhot, Tanya; Lavoie, Patrick; Cournoyer, Alexis; Balli, Fabio; Mathieu-Dupuis, Gabrielle

VERSION 1 – REVIEW

REVIEWER	Elizabeth A Boyle University of the West of Scotland, Scotland
REVIEW RETURNED	24-Oct-2017

GENERAL COMMENTS	Educational games have become very popular in recent years, largely because of how engaging and challenging entertainment games seem to be. A number of reviews of how effective games are in helping students to learn or change their behaviour have been carried out. The current review differs from previous reviews in proposing a specific focus on comparing the effects of different game design elements on engagement and performance. It is certainly true that there is a need to examine exactly which characteristics of games (design elements) lead to improvements in engagement and performance and such a study would certainly make a significant impact on games research. Given the journal the focus is of course on the use of games in a healthcare setting. The authors provide a clear account of how the study will be conducted and have completed the PRISMA checklist describing the protocol for this systematic review. They propose carefully considered search terms, and identify several different dimensions along which they will categorise studies. Outcome measures are well described. The authors also provide a clear account of the proposed analysis. However a number of clarifications would be useful. Types of participants Clarify that "all levels of education (from undergraduate to postgraduate education) either in a clinical (continuing education) or an academic setting" does not include children as there are health games directed towards children. Provide rationale for exclusion of health games for children. Types of interventions It is considered as good practice to include debriefing sessions when

	games are used to contextualise the use of the game, so provide further justification for only including “studies assessing a SG as a standalone intervention”. Types of comparators The authors provide a clear account of the DEs they have selected but they recognise that it may be difficult to find studies that specifically focus on “two groups receiving each a SG that varies for at least one DE between groups” and they propose to consider “studies where the comparator is any other type of educational intervention.” I wasn’t sure what this meant, as the focus is meant to be on games? Have the authors carried out a small pilot study to establish whether there are likely to be sufficient papers that report on the DEs as this seems crucial to the proposed review. While I ticked yes for “Is the standard of written English acceptable for publication?” the paper needs to be edited in places for clarity and correctness of English expression.

REVIEWER	Ann DeSmet Ghent University, Belgium
REVIEW RETURNED	13-Nov-2017

GENERAL COMMENTS	This paper describes a protocol for a review and meta-analysis of serious games for health professional education. The topic is interesting and the results would be a useful addition to literature in this field. My main concern with this paper is that the methods and decisions for the meta-analysis are very vague. The authors state that all will depend on the quality of the data. I would argue that the authors can impose restrictions in their search on the quality of the papers (e.g. only include RCTs, that have used validated scales to measure the outcome) to be included in the analysis. In my opinion, as a protocol for a meta-analysis, it is too weak. Many questions remain unanswered, e.g. which condition to include if more than two interventions are compared using the same control condition; what if certain values are reported adjusting for influence of covariates whereas others did not; what is the minimum number of participants per arm needed to be included,… . These are all crucial elements in a protocol for a meta-analysis. I believe this paper is therefore acceptable as a protocol to describe a systematic review, but not to describe the meta-analysis. I would recommend that authors either decide to only focus on the systematic review and describe a detailed meta-analysis protocol with the paper showing their meta-analytic results; or to substantially revise this paper and include all main decisions for the meta-analysis. Detailed comments: - introduction, p7: engagement has also been negatively related to learning outcomes, where a game was highly immersive and increased cognitive load distracting from the educational content (see Schrader & Bastiaens 2012). This nuance should be included here - introduction, p7: as design elements, only different forms of rewards have been mentioned. This is a very narrow definition of a game, fitting more with a gamification approach than with a serious game. I would suggest updating this list of examples to better reflect

	the main properties of a game - introduction, p8: "recent theoretical papers about SG development reported that the optimal integration of DEs has yet to be found". Whereas this may be true for serious games in healthcare professional education, such papers and meta-analyses do, however, exist for serious health games in other areas than professional development. As it may be assumed that the role of DEs can be similar across applications, I would recommend that the authors refer to findings from other serious health games as a foundation for hypotheses (see e.g. several papers in Games for Health journal; meta-analysis of DeSmet et al. 2014 showing the importance of theory and tailoring in game effectiveness) - Methods section needs to updated to include decisions for meta-analyses if the authors choose to present it as a protocol for meta-analysis - Page 9: typo acceptability? - Page 9: definition of educational outcomes: do all three outcomes need to be measured to qualify for inclusion? or is one of the three also sufficient? - Page 9: please add Games for Health journal to your selection of specialized journals for your handsearching method - Page 11: please include statistical information on interrater reliability for your coding (e.g. ICC, kappa) - Page 12: what is meant by imbalance between groups? please clarify - Page 12: please provide examples for the outcomes, especially on Behavior change - Page 14: random effects does not depend on heterogeneity! The choice to use random effects model depends on how the studies have measured their outcomes. If there is much variation in measurement method, random effects should be used rather than fixed effects. Please consult the manual of Borenstein et al. on conducting a meta-analysis - Page 14: 'if we are unable to explain heterogeneity this way, the synthesis will be merely narrative'. This is a strange method. Within a category of games, there may still be heterogeneity. This merely means that there are further 'subcategories' not explained by the moderator already used to group the interventions. This does not disqualify the meta-analytic results that you obtain. Please clarify why you mean here and provide an established reference for this method

REVIEWER	Elisabeth Whyte Penn State University, United States
REVIEW RETURNED	14-Nov-2017

GENERAL COMMENTS	The current study registers the protocol for a systematic review and meta-analysis of serious games. The current protocol examines how design elements may influence learning and engagement from serious games. The protocol is clear and follows PRISMA-P guidelines. However, the authors should be aware of another recent review of interest by Wang, DeMaria, Goldberg, & Katz (2016), “A systematic review of serious games in training health care professionals” that may have important overlap in studies reviewed. The current study differs in the moderator analyses related to examining design elements that was not included in this recent analysis by Wang et al.

	(2016). In addition, the choice of Design Elements does not seem to be directed by psychological or educational theories of how these game design elements may support learning. Previous research has examined how design elements may influence learning in other learning domains. For example, Wouters, van Nimwegen, van Oostendorp, & van der Spek (2013) found that narrative did not have an impact on learning from serious games. However, they hypothesize that how well the narrative is aligned with the educational goals matters more than the simple categorization of narrative present/absent. This was explored more in the Clark, Tanner-Smith, & Killingsoworth (2016) review of serious games (for children) that explored design elements as moderators based on breaking down the design characteristics into categories that may be meaningful (such as the depth or relevance of the story characteristics). Without a clear theoretical motivation for the choice of design elements and the hypothesized mechanism of action, it is difficult to ensure that the meta-analysis will have actionable insight about choice of design elements. Thus, in interpreting the results of the meta-analysis, the authors should be careful about collapsing across potentially meaningful design choices related to the categories of design elements listed in Table 1.

VERSION 1 – AUTHOR RESPONSE

Response to editorial comments

Editorial comment #1
“Please add the PROSPERO reference number to the manuscript on pages 3, 6 and 14”.

Response to editorial comment #1
The PROSPERO reference number was added to the manuscript on corresponding pages. (p.4-5-7-15-16)

Editorial comment #2
“Please clarify in the Abstract >> 'Methods and analysis' section which databases will be searched”.

Response to editorial comment #2
The databases searched were added to the abstract section. (p.4)

Editorial comment #3
“Please add the month as well as year to the dates of coverage for your literature search (page 7)”.

Response to editorial comment #3
The month as well as year to the dates of coverage of our literature search were added. (p.8)

Response to the comments of Reviewer #1

Comment #1
“Types of participants. Clarify that “all levels of education (from undergraduate to postgraduate education) either in a clinical (continuing education) or an academic setting” does not include children

as there are health games directed towards children. Provide rationale for exclusion of health games for children".

Response to comment #1
Clarifications were added to make it clear that studies conducted exclusively among any group of patients or students receiving education not related to healthcare delivery were beyond the scope of this review. For instance, the objectives of the games evaluated in these populations (e.g. promoting medication adherence in patients or teaching mathematical concepts in high school students) differ from the educational ones of serious games developed for healthcare professionals and healthcare students (which will be in the domain of training and professional development). (p.8)

Comment #2
"It is considered as good practice to include debriefing sessions when games are used to contextualise the use of the game, so provide further justification for only including "studies assessing a SG as a standalone intervention".

Response to comment #2
We thank the reviewer for this idea. In the "Types of interventions" section, we have now included serious game either as a standalone intervention or as part of a multi-component intervention. In addition, to describe which combinations were studied, we will perform subgroup analysis to compare the effectiveness of serious game when it is delivered as a standalone intervention or as part of a multi-component intervention. (p.8)

Comment #3
Type of comparators. The authors provide a clear account of the DEs they have selected but they recognise that it may be difficult to find studies that specifically focus on "two groups receiving each a SG that varies for at least one DE between groups" and they propose to consider "studies where the comparator is any other type of educational intervention." I wasn't sure what this meant, as the focus is meant to be on games? Have the authors carried out a small pilot study to establish whether there are likely to be sufficient papers that report on the DEs as this seems crucial to the proposed review.

Response to comment #3".
The "Type of comparators" section has been reviewed. We reworded it to make explicit the fact that we will include studies where the comparator is any type of educational intervention. In the systematic review by Wang et al. (2016), nine (9) studies evaluating the effectiveness of a serious game (pre/post-test measure, use of a control group) were identified. As their literature search was completed in 2014, it is expected that the last three years saw a growth in published studies. Indeed, these authors noted in their review an acceleration in published literature starting the beginning of this decade. Furthermore, the New Media Consortium stated in 2014 that the use of serious games was to be significantly experimented by educational institutions in the next two to three years. (p.8)

Comment #4
"While I ticked yes for "Is the standard of written English acceptable for publication?" the paper needs to be edited in places for clarity and correctness of English expression".

Response to comment #4".
The paper has now been entirely reviewed for clarity and correctness of English expression.

Response to the comments of Reviewer #2

Comment #1
“My main concern with this paper is that the methods and decisions for the meta-analysis are very vague. The authors state that all will depend on the quality of the data. I would argue that the authors can impose restrictions in their search on the quality of the papers (e.g. only include RCTs, that have used validated scales to measure the outcome) to be included in the analysis. In my opinion, as a protocol for a meta-analysis, it is too weak. Many questions remain unanswered, e.g. which condition to include if more than two interventions are compared using the same control condition; what if certain values are reported adjusting for influence of covariates whereas others did not; what is the minimum number of participants per arm needed to be included,... . These are all crucial elements in a protocol for a meta-analysis. I believe this paper is therefore acceptable as a protocol to describe a systematic review, but not to describe the meta-analysis. I would recommend that authors either decide to only focus on the systematic review and describe a detailed meta-analysis protocol with the paper showing their meta-analytic results; or to substantially revise this paper and include all main decisions for the meta-analysis.”.

Response to comment #1

We agree with the reviewer that specific criteria needed to be explicitly stated. We have updated the eligibility criteria section and now clearly state that only randomised controlled trials (RCT) and cluster-RCT will be included in the systematic review and the meta-analysis, as recommended by the EPOC Cochrane Review Group.

For each stated outcome, we will prioritize adjusted data, when available. We will pool, for each specified outcome, all included studies reporting on that outcome (as it is anticipated that not each included study will not report on all specified outcomes) to get an overall effect size.

Furthermore, as recommended by Cochrane, we do not plan on prospectively excluding studies solely based on their sample size. In this sense, Cochrane does not specify a minimal number of participants needed per arm in order to be included in a meta-analysis. If there is evidence of statistical heterogeneity, we will compare the estimates of the effect size between fixed- and random-effect models. If the effect given by the random-effect appears superior, we will investigate if the exclusion of smaller studies from the meta-analysis has an impact on the finding. (p.12)

Comment #2
“introduction, p7: engagement has also been negatively related to learning outcomes, where a game was highly immersive and increased cognitive load distracting from the educational content (see Schrader & Bastiaens 2012). This nuance should be included here”

Response to comment #2
This nuance has now been included at the end of the third paragraph of the introduction (p.6-7).

Comment #3
“introduction, p7: as design elements, only different forms of rewards have been mentioned. This is a very narrow definition of a game, fitting more with a gamification approach than with a serious game. I would suggest updating this list of examples to better reflect the main properties of a game”

Response to comment #3
The list of examples was updated to better reflect the diversity of design elements to be assessed in this systematic review. The full list of elements is presented in Table 1. (p.6; p.21-22)

Comment #4

“introduction, p8: "recent theoretical papers about SG development reported that the optimal integration of DEs has yet to be found". Whereas this may be true for serious games in healthcare professional education, such papers and meta-analyses do, however, exist for serious health games in other areas than professional development. As it may be assumed that the role of DEs can be similar across applications, I would recommend that the authors refer to findings from other serious health games as a foundation for hypotheses (see e.g. several papers in Games for Health journal; meta-analysis of DeSmet et al. 2014 showing the importance of theory and tailoring in game effectiveness)”

Response to comment #4
We have updated this section and now refer to the findings, among others, of DeSmet et al. 2014. (p.7)

Comment #5
“Methods section needs to updated to include decisions for meta-analyses if the authors choose to present it as a protocol for meta-analysis”

Response to comment #5
We have now updated this section to make explicit the decisions for conducting a meta-analysis. (p.8; p.13)

Comment #6
“Page 9: typo acceptability?”

Response to comment #6
We corrected this typo. (p.9)

Comment #7
“Page 9: definition of educational outcomes: do all three outcomes need to be measured to qualify for inclusion? or is one of the three also sufficient?”

Response to comment #7
No. The section “Type of outcome measures” was clarified to make it explicit that studies reporting at least one measure of engagement or educational outcome will be considered for inclusion. (p.9)

Comment #8
“Page 9: please add Games for Health journal to your selection of specialized journals for your handsearching method”

Response to comment #8
As this journal also focuses on clinical training, it was added to our selection. (p.10)

Comment #9
“Page 11: please include statistical information on interrater reliability for your coding (e.g. ICC, kappa)”

Response to comment #9
We thank the reviewer for this suggestion. The Cochrane Handbook of Systematic Review of Interventions doesn’t recommend to routinely quantify agreement of coded items and we had not planned to do so. However, his comment made us realize the need to pilot our form to ensure its understanding by the two reviewers who will extract the data. A clarification about this has been

added to the protocol. Furthermore, we will make sure to resolve any discrepancies between the forms completed by the two reviewers before inputting the data. (p.11)

Comment #10
“Page 12: what is meant by imbalance between groups? please clarify”

Response to comment #10
We clarified this section. It now reads: “statistical differences at baseline between groups”. (p.11)

Comment #11
“Page 12: please provide examples for the outcomes, especially on Behavior change”

Response to comment #11
Examples for each outcome are now provided in the “Type of outcome measures”. (p.9)

Comment #12
“Page 14: random effects does not depend on heterogeneity! The choice to use random effects model depends on how the studies have measured their outcomes. If there is much variation in measurement method, random effects should be used rather than fixed effects. Please consult the manual of Borenstein et al. on conducting a meta-analysis”

Response to comment #12
Again, we thank the reviewer for her remark. We have now reviewed the justification for using a random-effect model in our planned meta-analysis. (p.13)

Comment #13
Page 14: 'if we are unable to explain heterogeneity this way, the synthesis will be merely narrative'. This is a strange method. Within a category of games, there may still be heterogeneity. This merely means that there are further 'subcategories' not explained by the moderator already used to group the interventions. This does not disqualify the meta-analytic results that you obtain. Please clarify why you mean here and provide an established reference for this method

Response to comment #13
As recommended by the Cochrane Handbook of Systematic Reviews of Interventions, in the case of important heterogeneity, we will investigate the source of heterogeneity by conducting the planned sensitivity analysis and subgroup analysis. However, if we are unable to address heterogeneity, Cochrane suggests to not perform a meta-analysis as it may be misleading to quote an average value for the intervention effect. (p.14-15)

Response to the comments of Reviewer #3

Comment #1
However, the authors should be aware of another recent review of interest by Wang, DeMaria, Goldberg, & Katz (2016), “A systematic review of serious games in training health care professionals” that may have important overlap in studies reviewed. The current study differs in the moderator analyses related to examining design elements that was not included in this recent analysis by Wang et al. (2016).

Response to comment #1

Indeed, we are aware of this previous systematic review. As the reviewer correctly pointed it, our review will first differ from the sensitivity analyses that are planned to examine the impact of design elements. Second, as the review by Wang et al. (2016) reviewed the literature until the end of the 2014, our review will also add by synthesizing the literature published in the last three (3) years Our systematic review will also add to the literature by assessing and reporting bias using a validated assessment tool. We also plan on conducting a meta-analysis, something which had not been done in their review or in any other review with a specific focus on training healthcare professionals and healthcare students, according to our knowledge.

Comment #2
In addition, the choice of Design Elements does not seem to be directed by psychological or educational theories of how these game design elements may support learning. Previous research has examined how design elements may influence learning in other learning domains. For example, Wouters, van Nimwegen, van Oostendorp, & van der Spek (2013) found that narrative did not have an impact on learning from serious games. However, they hypothesize that how well the narrative is aligned with the educational goals matters more than the simple categorization of narrative present/absent. This was explored more in the Clark, Tanner-Smith, & Killingsoworth (2016) review of serious games (for children) that explored design elements as moderators based on breaking down the design characteristics into categories that may be meaningful (such as the depth or relevance of the story characteristics). Without a clear theoretical motivation for the choice of design elements and the hypothesized mechanism of action, it is difficult to ensure that the meta-analysis will have actionable insight about choice of design elements. Thus, in interpreting the results of the meta-analysis, the authors should be careful about collapsing across potentially meaningful design choices related to the categories of design elements listed in Table 1.

Response to comment #2

We wish to begin by thanking the reviewer for her comment. From the outset, the choice of design elements was based on a review of the literature to identify the elements most frequently found in serious games today. We currently hypothesize that these design elements operate by influencing one of the engagement's antecedents, as currently described at the end of the 2nd paragraph of the introduction. These antecedents of engagement were initially proposed in the flow theory of Csikszentmihalyi (1990), a theory that has served in the development of models or theoretical propositions dealing more specifically with the influence of design elements in video games on the players' engagement. Pavlas (2010), as part of his doctoral dissertation on the impact of certain video game characteristics on engagement and learning, also took up these models and theoretical propositions. Based on our examination of the literature, we suggest that the retained design elements are related to an antecedent of engagement, and that engagement will be related to educational outcomes.

In response to the commentary of the reviewer we have expanded the theoretical explanations in the introduction to make more explicit the proposed mechanism of action of these design elements. In the data to be collected, we were already planning on collecting the theoretical frameworks on which the authors based the development of their serious games. We will examine if the choice of the design elements made by the authors were related to the theoretical foundations they present. Finally, we made explicit the limitation to our review regarding the missing/present dichotomization. However, we will take care to contrast the results that we will obtain with the results of the previous reviews stated by the reviewer. Nonetheless, we will still be able to make recommendations for the development of serious games with our study population regarding the design elements that have received the least attention and how the current elements are integrated.

VERSION 2 – REVIEW

REVIEWER	Elisabeth Whyte Penn State, United States
REVIEW RETURNED	26-Dec-2017

GENERAL COMMENTS	The authors have been responsive to the concerns raised in the previous review.

REVIEWER	Ann DeSmet Ghent University, Belgium
REVIEW RETURNED	03-Jan-2018

GENERAL COMMENTS

The paper has greatly improved, but some major comments remain regarding the methodology of the meta-analysis. These are listed below.

* Author comment “Examples for each outcome are now provided in the “Type of outcome measures”. (p.9)”
The authors state that they will also investigate effectiveness on engagement? (page 14)? It is not yet included in the protocol how these outcomes will be measured. A reference to another paper is also not sufficient here. What is needed, is to know: how will you standardize outcomes across these levels of engagement? Will you include all types of engagement and consider them as equivalent, or will you investigate e.g. drop-out and adherence as one type of engagement outcome, and subjective experience of fun as another type of engagement outcome? The protocol is currently too vague on this topic and needs further elaboration if the purpose is to indeed conduct a meta-analysis on this outcome.

* Page 15 ‘If sufficient data are available, we will try to explain the source of statistical heterogeneity by exploring clinical and methodological diversity’.
How will the authors decide what qualifies as sufficient data? Please read the paper by Susanne Hempel et al, in Systematic Reviews, 2013 on the topic of power analyses for moderator analyses in meta-analysis, and include a decision in your protocol. How you will make the decision on whether the moderator analysis can be carried out with sufficient power, based on the number of studies or combined sample across studies for each moderator, should be included in your protocol.
The authors have included more information on moderator and sensitivity analyses, but I found the description on items to be coded on page 13 confusing. I assumed these would be included as moderators, but then see on page 15, that this is not the case, and that these are items to be used as descriptive variables in the systematic review. Please make it clear in the text that these are intended for the systematic review, and not the meta-analysis.
I moreover feel that the list of moderators on page 15 (referred to as subgroup analysis) for the meta-analysis is too limited. What should be added here is:
- Type of outcome: knowledge, behavior change, type of engagement outcome
- The specific design elements: this is listed in the title of the paper (impact of design elements on engagement and effectiveness), yet these do not figure in the list of moderators?

	- Theoretical framework There are quite a few of items mentioned under data extraction, of which is currently not clear whether these will also be included in the moderator analyses. At the very least, the authors should state whether they will attempt to analyze the impact of these features on effectiveness, where the feasibility during the study will depend on whether there are sufficient observations to run these with enough power (see first comment) Also with respect to the sensitivity analyses, I think certain elements are missing: - Outlier analysis: please check whether any of your outcomes can be considered an outlier and remove this. This would greatly affect your data - Mention quality of the studies in sensitivity analyses. You may want to consider only higher quality studies in the moderator analyses if a significant difference is noticed - What about adjusted ratios compared to raw figures? * "Page 8: For this systematic review, we define SGs as interactive digital software with a primary educational purpose that engage the a learner through various challenges." I would encourage the authors to include the element of aiming to entertain /be fun in their definition for serious games. The authors state there is no clear definition of serious games, but SG's aim to be both educational and fun is quite widespread as definition in serious game literature. The current definition 'of engaging via challenges' would also include a digital tool that asks some questions after the educational part and where the person may receive a score out of 10, but which can hardly be considered a game since it does not intend to be entertaining or fun. * Authors' comment "We thank the reviewer for this suggestion. The Cochrane Handbook of Systematic Review of Interventions doesn't recommend to routinely quantify agreement of coded items and we had not planned to do so. However, his comment made us realize the need to pilot our form to ensure its understanding by the two reviewers who will extract the data. A clarification about this has been added to the protocol. Furthermore, we will make sure to resolve any discrepancies between the forms completed by the two reviewers before inputting the data. (p.11)" I think the authors need to be wary that game design elements may be harder to code than some rather obvious study characteristics such as sample size and study duration. This is a qualitative analysis, rather than a mere data extraction of undisputable items. Not showing interrater reliability and checking the finalized checklist of some coded examples with a few experts, would seriously reduce validity and reliability of the results. I for one would never cite papers where validity and reliability have not been demonstrated, regardless of how advanced the statistics were on the topic. It's the basis of everything that follows, if the validity and reliability of coding is not demonstrated, the results have no value. Please also mention whether you will code these based on the information in the paper, of based on actual game play. Some further comments on the table of design elements: I would not consider points as immediate feedback, in line with psychological definitions of the concept of immediate feedback. Story: 'a narrative context', I do not feel this is the correct conceptualization of a story. E.g. a story needs to have a plot. Please refer to papers written by Amy Lu Shirong for the use of stories in serious health games

	There are still some grammatical mistakes in the table of design elements. * "If multiple comparisons are relevant in a single study, we will split the "shared" group in multiple subgroups to allow pair-wise comparisons." The problem here is double counting the control condition. For example, if you have three intervention groups and one control condition, with each n=20, taking the three intervention groups as multiple subgroups in the meta-analysis means that you are counting your control condition as n=60 instead of n=20 (since it is being included three times in the meta-analysis), doing so affects the influence of this study on the pooled effect size and the confidence intervals. If you have several intervention groups in one study, each with their own control group, then it is ok to proceed as you suggested. If only one control group is used to compare several intervention groups against, you cannot combine these in one overall analysis. This may be what you refer to in the 'splitting shared group in multiple subgroups to allow pair-wise comparisons', but this is not sufficiently clear. I would suggest to phrase it as "when multiple subgroup comparisons were made against a single control group, these effects are only reported in separate subgroup analyses and not combined in one overall effect size calculation to avoid double counting of sample size of the control group". However, you may still want to decide what to do with this study when calculating your overall effect size, and on how you will decide which one of these three intervention groups would be most relevant for that. My suggestion would be that you consider which of these intervention groups is the least represented to have a broader overview of SG's, or you may want to use the intervention group that fits the most of your coded characteristics to allow more moderator analyses (see power analyses in an earlier comment) * Authors' comment "Again, we thank the reviewer for her remark. We have now reviewed the justification for using a random-effect model in our planned meta-analysis. (p.13): The decision to use random-effect models was made due to the expected variability between SGs (notably on DEs)." Please include 'variability in study design', to clarify that you are not merely looking at heterogeneity to decide whether or not to use random effects * "Page 18: based on the SG type (e.g. quiz, management)." What is 'management' as a type of serious game? Please refer to some established literature on game genres (e.g. simulation games, role-playing games, mystery) instead

VERSION 2 – AUTHOR RESPONSE

Response to the comments of Reviewer #2

Comment #1
Author comment "Examples for each outcome are now provided in the "Type of outcome measures". (p.9)"
The authors state that they will also investigate effectiveness on engagement? (page 14)? It is not yet included in the protocol how these outcomes will be measured. A reference to another paper is also not sufficient here. What is needed, is to know: how will you standardize outcomes across these

levels of engagement? (#1) Will you include all types of engagement and consider them as equivalent, or will you investigate e.g. drop-out and adherence as one type of engagement outcome, and subjective experience of fun as another type of engagement outcome? (#2) The protocol is currently too vague on this topic and needs further elaboration if the purpose is to indeed conduct a meta-analysis on this outcome.

Response to comment #1
Indeed, it is currently planned to investigate the effectiveness of serious games and the impact of design elements on engagement. We retain the definition proposed by Perski et al. (2017) in their systematic review of the literature as a two-dimensional concept (i.e. 1) extent of the learner's involvement and 2) a subjective experience characterised by affect, attention, and interest). These two dimensions will be considered individually in the planned review and meta-analysis. Considering involvement, we will look individually at two variables: duration and frequency of serious game usage. Considering the subjective experience, we will look at self-reported measures of the learner's experience while using the SG. As it is anticipated that the variables related to this dimension will vary (e.g. interest, fun) throughout the studies, we aren't currently planning on pooling together all variables related to this dimension regardless of what they are or the definition given to them by the authors of the original studies. When the variables related to this specific dimension will be extracted, we will look at whether or not it makes sense to pool them together and if they allow the performing of a meta-analysis. In all cases, we will describe which variables related to this dimension were considered in the original studies.

In response to the reviewer comment, we added specifications in the "Types of outcome measures" (P9L6-12) and in the "Data Items" sections (P12L14-16).

Comment #2
* Page 15 'If sufficient data are available, we will try to explain the source of statistical heterogeneity by exploring clinical and methodological diversity'.
How will the authors decide what qualifies as sufficient data? (#1) Please read the paper by Susanne Hempel et al, in Systematic Reviews, 2013 on the topic of power analyses for moderator analyses in meta-analysis, and include a decision in your protocol. How you will make the decision on whether the moderator analysis (#2) can be carried out with sufficient power, based on the number of studies or combined sample across studies for each moderator, should be included in your protocol.

Response to comment #2
We plan to conduct subgroup analysis when there are at least two studies that can be included in one planned subgroup analysis. However, we want to stress that we currently plan to perform subgroup analysis primarily to investigate statistical heterogeneity and not to identify a possible "genuine" difference in the magnitude of effect between subgroups. Caution in the interpretation of subgroup analyses is underlined in the Cochrane Handbook of Systematic Reviews of Interventions as they remain observational. Therefore, we are not currently planning on performing power calculation in order to perform these subgroup analyses. Furthermore, no meta-regression is currently planned. If differences in the magnitude of effect are found between subgroup, their observational nature will be stressed when reported.

In response to the reviewer comment, we added specifications about the criteria for performing subgroup analyses (P14L25-27) and about the interpretation of subgroup analyses (P14L22-27).

Comment #3
The authors have included more information on moderator and sensitivity analyses, but I found the description on items to be coded on page 13 confusing. I assumed these would be included as

moderators, but then see on page 15, that this is not the case, and that these are items to be used as descriptive variables in the systematic review. Please make it clear in the text that these are intended for the systematic review, and not the meta-analysis.

Response to comment #3
We made sure to distinguish, in the “Data Items” section, between descriptive data and variables which will serve for the meta-analysis (P12L2-17).

Comment #4
I moreover feel that the list of moderators on page 15 (referred to as subgroup analysis) for the meta-analysis is too limited. What should be added here is:
- Type of outcome: knowledge, behavior change, type of engagement outcome
- The specific design elements: this is listed in the title of the paper (impact of design elements on engagement and effectiveness), yet these do not figure in the list of moderators?
- Theoretical framework
There are quite a few of items mentioned under data extraction, of which is currently not clear whether these will also be included in the moderator analyses. At the very least, the authors should state whether they will attempt to analyze the impact of these features on effectiveness, where the feasibility during the study will depend on whether there are sufficient observations to run these with enough power (see first comment)

Response to comment #4
As stated at the beginning of the “Quantitative data synthesis” section, we plan to perform a meta-analysis to evaluate the overall effect of serious games on each outcome stated in the “Types of outcome measures” section (i.e. a meta-analysis for each outcome). It is planned to evaluate the impact of design elements through sensitivity analysis (P15L11-13). Regarding the “theoretical framework” data item, it will be extracted for descriptive purpose and it is not planned to use it in the meta-analysis as it is anticipated that the specific theoretical framework used and its alignment with the serious game evaluated will greatly vary among studies. Furthermore, as the Cochrane Handbook of Systematic Reviews of Interventions recommends keeping subgroup analysis to a minimum in order to avoid the likelihood of false positives, we weren’t planning on adding more to the ones already present.

Comment #5
- Outlier analysis: please check whether any of your outcomes can be considered an outlier and remove this. This would greatly affect your data

Response to comment #5
We expect outlier results or extreme values to be associated with small studies or ones at high risk of bias. The impact of small studies on the results will be explored through sensitivity analysis (P15L9-11). High risk of bias studies will not be included in the meta-analysis (P13L21-22). However, we aren’t planning on excluding, for example, a large and well-conducted study on the basis that extreme values would be introduced unless a rationale can support this decision. Nonetheless, we’ll make sure to discuss the impact that this kind of study could have had on the results presented.

Comment #6
Mention quality of the studies in sensitivity analyses. You may want to consider only higher quality studies in the moderator analyses if a significant difference is noticed

Response to comment #6

It is currently planned to include only RCT and cluster-RCT in this systematic review and meta-analysis. “Quality of studies”, or risk of bias, will be assessed using the EPOC criteria. It is now planned to restrict the meta-analysis to only studies at low risk of bias. Incorporating studies at high or unclear risk of bias in the meta-analysis and discussing their impact on the results could still lead to a misinterpretation of the findings by the reader and decision-taking based on flawed evidence. Therefore, as recommended by the Cochrane Handbook of Systematic Reviews of Interventions, we will present narratively how the results might have been affected by “high risk of bias studies” but we will not formally and graphically present these results in the result section.

We added precisions in the “Quantitative data synthesis” section to stress the fact that the meat-analysis will be restricted to “low risk of bias studies” (P13L21-22).

Comment #7
What about adjusted ratios compared to raw figures?

Response to comment #7
When data adjusted for baseline differences between groups are available, we’ll use it to compute effect size. When adjusted data are not available, we’ll use unadjusted data.

We added a specification about the use of adjusted data in the “Data extraction process” section (P11L21-22).

Comment #8
* “Page 8: For this systematic review, we define SGs as interactive digital software with a primary educational purpose that engage the learner through various challenges.”
I would encourage the authors to include the element of aiming to entertain /be fun in their definition for serious games. The authors state there is no clear definition of serious games, but SG's aim to be both educational and fun is quite widespread as definition in serious game literature. The current definition ‘of engaging via challenges’ would also include a digital tool that asks some questions after the educational part and where the person may receive a score out of 10, but which can hardly be considered a game since it does not intend to be entertaining or fun.

Response to comment #8
The entertainment aspect is, indeed, an important aspect of SGs. We have reworked our definition to also state this (P4L2; P6L12; P8L23).

Comment #9
* Authors’ comment “We thank the reviewer for this suggestion. The Cochrane Handbook of Systematic Review of Interventions doesn’t recommend to routinely quantify agreement of coded items and we had not planned to do so. However, his comment made us realize the need to pilot our form to ensure its understanding by the two reviewers who will extract the data. A clarification about this has been added to the protocol. Furthermore, we will make sure to resolve any discrepancies between the forms completed by the two reviewers before inputting the data. (p.11)”
I think the authors need to be wary that game design elements may be harder to code than some rather obvious study characteristics such as sample size and study duration. This is a qualitative analysis, rather than a mere data extraction of undisputable items. Not showing interrater reliability and checking the finalized checklist of some coded examples with a few experts, would seriously reduce validity and reliability of the results. I for one would never cite papers where validity and reliability have not been demonstrated, regardless of how advanced the statistics were on the topic.

It’s the basis of everything that follows, if the validity and reliability of coding is not demonstrated, the results have no value.
Please also mention whether you will code these based on the information in the paper, of based on actual game play.

Response to comment #9
Indeed, as stated by the reviewer, the coding of design elements could be a challenging element. As such, we added precisions about how we plan to perform this step. Similarly to the other data items to be extracted, two review authors will independently perform the extraction of data items related to design elements (P11L9-11). Whenever, the serious game evaluated is publicly available, we’ll favor actual game play to code the design elements. When it is not possible, we will code the design elements based on the information provided in the original article (P11L16-18). All procedures regarding coding either based on the game play or the original article will be detailed in the results paper of this systematic review. We will also calculate a Kappa statistic to illustrate agreement on the extraction of this specific data item (P11L18-20).

Comment #10
Some further comments on the table of design elements:
I would not consider points as immediate feedback, in line with psychological definitions of the concept of immediate feedback.
Story: ‘a narrative context’, I do not feel this is the correct conceptualization of a story. E.g. a story needs to have a plot. Please refer to papers written by Amy Lu Shirong for the use of stories in serious health games
There are still some grammatical mistakes in the table of design elements.

Response to comment #10
We reviewed the Table 1 in order for the definitions provided to better reflect the current state of the literature and to correct grammatical mistakes (P21).

Comment #11
* “If multiple comparisons are relevant in a single study, we will split the “shared” group in multiple subgroups to allow pair-wise comparisons.”
The problem here is double counting the control condition. For example, if you have three intervention groups and one control condition, with each n=20, taking the three intervention groups as multiple subgroups in the meta-analysis means that you are counting your control condition as n=60 instead of n=20 (since it is being included three times in the meta-analysis), doing so affects the influence of this study on the pooled effect size and the confidence intervals. If you have several intervention groups in one study, each with their own control group, then it is ok to proceed as you suggested. If only one control group is used to compare several intervention groups against, you cannot combine these in one overall analysis. This may be what you refer to in the ‘splitting shared group in multiple subgroups to allow pair-wise comparisons’, but this is not sufficiently clear. I would suggest to phrase it as “when multiple subgroup comparisons were made against a single control group, these effects are only reported in separate subgroup analyses and not combined in one overall effect size calculation to avoid double counting of sample size of the control group”. However, you may still want to decide what to do with this study when calculating your overall effect size, and on how you will decide which one of these three intervention groups would be most relevant for that. My suggestion would be that you consider which of these intervention groups is the least represented to have a broader overview of SG’s, or you may want to use the intervention group that fits the most of your coded characteristics to allow more moderator analyses (see power analyses in an earlier comment)

Response to comment #11

We thank the reviewer for her comment and her recommendations. Based on the The Cochrane Handbook of Systematic Review of Interventions, we added further specifications regarding the handling of multiple-arm trials in the planned meta-analysis. First, only relevant pair-wise comparisons will be included and based on the “Eligibility criteria” section. Therefore, if a single study contributed to several independent comparisons (i.e. no group in common between comparisons), we’ll include all relevant comparisons. If one or more groups are shared between comparisons, we’ll first try to combine several groups in order to create a single pair-wise comparison. If it’s not possible, we’ll split the participants in the “shared group” into multiple groups, with smaller sample size, to allow pair-wise comparisons and to avoid double-counting (P13L25-28;P14L1-5). As stated in the Cochrane Handbook, the exclusion of relevant comparisons would lead to a loss of information.

Comment #12
* Authors’ comment “Again, we thank the reviewer for her remark. We have now reviewed the justification for using a random-effect model in our planned meta-analysis. (p.13): The decision to use random-effect models was made due to the expected variability between SGs (notably on DEs).”
Please include ‘variability in study design’, to clarify that you are not merely looking at heterogeneity to decide whether or not to use random effects

Response to comment #12
We have now included this precision in the article (P14L8-9).

Comment #13
* “Page 18: based on the SG type (e.g. quiz, management).”
What is ‘management’ as a type of serious game? Please refer to some established literature on game genres (e.g. simulation games, role-playing games, mystery) instead

Response to comment #13
Management would refer to games where the learner needs to manage resources in order to achieve in-game objectives; eMedOffice (Hannig et al., 2012) would be a serious game falling under that umbrella. We updated the examples in the list given in the article in order to avoid confusion among the reader. We also now refer to the game genre list proposed by Wang et al. (2016) (based originally on a work by Wolf (2015)) (P16L25).